Incrementally building FAIR Digital Objects with Specimen Data Refinery workflows

Oliver Woolland[‡], Paul Brack[‡], Stian Soiland-Reyes[‡], Ben Scott[§], Laurence Livermore[§]

‡ The University of Manchester, Manchester, United Kingdom
§ The Natural History Museum, London, United Kingdom

Corresponding author: Stian Soiland-Reyes (soiland-reyes@manchester.ac.uk), Laurence Livermore (l.livermore@nhm.ac.uk)

Abstract

Specimen Data Refinery (SDR) is a developing platform for automating transcription of specimens from natural history collections (Hardisty et al. 2022). SDR is based on computational workflows and digital twins using FAIR Digital Objects.

We show our recent experiences with building SDR using the Galaxy workflow system and combining two FDO methodologies with open digital specimens (openDS) and RO-Crate data packaging. We suggest FDO improvements for incremental building of digital objects in computational workflows.

SDR workflows

SDR is realised as the workflow system Galaxy (Afgan et al. 2018) with SDR tools installed. An Open Research challenge is that some tools have machine learning models with a commercial licence. This complicates publishing to Galaxy toolshed, however we created Ansible scripts to install equivalent Galaxy servers, including tools and dependencies, accounts and workflows. SDR workflows are published in WorkflowHub as FDOs.

We implemented the use case *De novo digitization* in Galaxy (Brack et al. 2022). Shown in Fig. 1 the workflow steps exchange openDS JSON (Hardisty et al. 2019), for incremental completion of a digital specimen. Initial stages build a template openDS from a CSV with metadata and image references – subsequent analysis completes the rest of the JSON with *regions* of interest, *text* digitised from handwriting, and recognized *named entities*.

Galaxy can visualise outputs of each step (Fig. 2), important to make the FDOs understandable by domain experts and to verify accuracy of SDR.

We are adding workflows for partial stages, e.g. detection of regions (Livermore and Woolland 2022a) and hand-written text recognition (Livermore and Woolland 2022b), which we'll combine with scalability testing and wider testing by project users. Additional

workflows will enhance existing FDOs and use new tools such as barcode detection of museums' internal identifiers.

We are now ready to publish digital specimens as FAIR Digital Objects, with registration into DiSSCO repositories, PID assignment and workflow provenance. However, even at this early stage we have identified several challenges that need to be addressed.

FDO lessons

We highlight the *De novo* use case because this workflow is exchanging *partial* FDOs – openDS objects which are not fully completed and not yet assigned persistent identifiers. openDS schemas are still in development, therefore SDR uses a more flexible JSON schema where only the initial metadata (populated from CSV) are required. Each step validates the partial FDO before passing it to the underlying command line tool.

Although workflow steps exchange openDS objects, they cannot be combined in any order. For instance, *named entity recognition* requires digitised text in the FDO. We can consider these intermediate steps as *sub-profiles* of an FDO Type. Unlike hierarchical subclasses, these FDO profiles are more like ducktyping. For instance a *text detection* step may only require the *regions* key, but semantically there is no requirement for an *OpenDSWithText* to be a subclass of *OpenDSWithRegion*, as text also can be transcribed manually without regions.

Similarly, we found that some steps can be executed in parallel, but this requires merging of partial FDOs. This can be achieved by combining JSON queries and JSON Schemas, but indicates that it may be more beneficial to have FDO fragments as separate objects. Adding openDS fragment steps would however complicate workflows.

Several of our tools process the referenced images, currently https URLs in openDS. We added a caching layer to avoid repeated image downloading, coupled with local file-paths wiring in the workflow. A similar challenge occurs if accessing image data using DOIP, which unlike HTTP, has no caching mechanisms.

RO-Crate lessons

Galaxy is developing support for importing and exporting Workflow Run Crates, a profile of RO-Crate (Soiland-Reyes et al. 2022b) to captures execution history of a workflow, including its definition and intermediate data (De Geest et al. 2022). SDR is adopting this support to combine openDS FDOs with workflow provenance, as envisioned by Walton et al. (2020).

Our prototype *de novo* workflow returns results as a ZIP file of openDS objects. End-users should also get copies of the referenced images and generated visualisations, along with workflow execution metadata. We are investigating ways to embed the preliminary Galaxy workflow history before the final step, so that this result can be an enriched RO-Crate.

Conclusions

SDR is an example of machine-assisted construction of FDOs, which highlight the needs for intermediate digital objects that are not yet FDO compliant. The passing of such “local FDOs” is beneficial not just for efficiency and visual inspection, but also to simplify workflow composition of canonical workflow building blocks. At the same time we see that it is insufficient to only pass FDOs as JSON objects, as they also have references to other data such as images, which should not need to be re-downloaded.

Further work will investigate the use of RO-Crate as a wrapper of partial FDOs, but this needs to be coupled with more flexible FDO types as profiles, in order to restrict “impossible” ordering of steps depending on particular inner FDO fragments. A distinction needs to be made between open digital specimens that are in “draft” state and those that can be pushed to DiSSCo registries.

We are experimenting with changing the SDR components into Canonical Workflow Building Blocks (Soiland-Reyes et al. 2022a) using the Common Workflow Language (Crusoe et al. 2022). This gives flexibility to scalably execute SDR workflows on different compute backends such as HPC or local cluster, without the additional setup of Galaxy servers.

Keywords

FDO, research object, RO-Crate, computational workflow, Galaxy, openDS, specimen, digitization

Presenting author

Stian Soiland-Reyes

Presented at

First International Conference on FAIR Digital Objects, poster

Acknowledgements

We acknowledge the SYNTHESYS+ and DiSSCO project members who have been invaluable in early evaluation and feedback on the development of SDR.

Author contributions

Author contributions to this article according to the Contributor Roles Taxonomy CASRAI CrEDiT:

- **Oliver Woolland**: Data curation, Resources, Software, Visualization, Writing – review & editing
- **Paul Brack**: Conceptualization, Software
- **Stian Soiland-Reyes**: Investigation, Methodology, Supervision, Writing – original draft, Writing – review & editing
- **Ben Scott**: Data curation, Software, Validation
- **Laurence Livermore**: Conceptualization, Data curation, Funding acquisition, Methodology, Project administration, Resources, Writing – review & editing

Conflicts of interest

References

- Afgan E, Baker D, Batut B, van den Beek M, Bouvier D, Čech M, Chilton J, Clements D, Coraor N, Grüning BA, Guerler A, Hillman-Jackson J, Hiltemann S, Jalili V, Rasche H, Soranzo N, Goecks J, Taylor J, Nekrutenko A, Blankenberg D (2018) The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic Acids Research 46 https://doi.org/10.1093/nar/gky379
- Brack P, Woolland O, Livermore L (2022) De novo digitisation. WorkflowHub https://doi.org/10.48546/workflowhub.workflow.373.1
- Crusoe M, Abeln S, Iosup A, Amstutz P, Chilton J, Tijanić N, Ménager H, Soiland-Reyes S, Gavrilović B, Goble C, Community TC (2022) Methods included. Communications of the ACM 65 (6): 54-63. https://doi.org/10.1145/3486897
- De Geest P, Coppens F, Soiland-Reyes S, Eguinoa I, Leo S (2022) Enhancing RDM in Galaxy by integrating RO-Crate. (submitted).
- Hardisty A, Ma K, Nelson G, Fortes J (2019) 'openDS' – A New Standard for Digital Specimens and Other Natural Science Digital Object Types. Biodiversity Information Science and Standards 3 https://doi.org/10.3897/biss.3.37033
- Hardisty A, Brack P, Goble C, Livermore L, Scott B, Groom Q, Owen S, Soiland-Reyes S (2022) The Specimen Data Refinery: A Canonical Workflow Framework and FAIR Digital Object Approach to Speeding up Digital Mobilisation of Natural History Collections. Data Intelligence 4 (2): 320-341. https://doi.org/10.1162/dint_a_00134
- Livermore L, Woolland O (2022a) DLA-Collections-test. WorkflowHub https://doi.org/10.48546/workflowhub.workflow.374.1
- Livermore L, Woolland O (2022b) HTR-Collections-test. WorkflowHub https://doi.org/10.48546/workflowhub.workflow.375.1

- Soiland-Reyes S, Bayarri G, Andrio P, Long R, Lowe D, Niewielska A, Hospital A, Groth P (2022a) Making Canonical Workflow Building Blocks Interoperable across Workflow Languages. Data Intelligence 4 (2): 342-357. https://doi.org/10.1162/dint_a_00135
- Soiland-Reyes S, Sefton P, Crosas M, Castro LJ, Coppens F, Fernández J, Garijo D, Grüning B, La Rosa M, Leo S, Ó Carragáin E, Portier M, Trisovic A, RO-Crate Community, Groth P, Goble C (2022b) Packaging research artefacts with RO-Crate. Data Science 5 (2). https://doi.org/10.3233/ds-210053
- Walton S, Livermore L, Bánki O, Cubey R, Drinkwater R, Englund M, Goble C, Groom Q, Kermorvant C, Rey I, Santos C, Scott B, Williams A, Wu Z (2020) Landscape Analysis for the Specimen Data Refinery. Research Ideas and Outcomes 6 https://doi.org/10.3897/rio.6.e57602

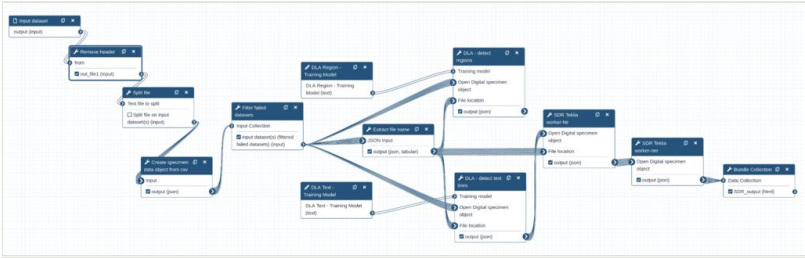

Figure 1.

Draft Galaxy workflow *De Novo digitization* (Brack et al. 2022) shows propagation of partial Open Digital Specimen FDOs between individual canonical workflow building blocks. First steps process a CSV file to create the initial openDS, where referenced images are analysed to detect text lines which are OCRed and then recognized as named entities. Bands indicate flow of collections of openDS, processed concurrently by each step. The final step bundles the collection of openDS FDOs as JSON files in a ZIP archive.

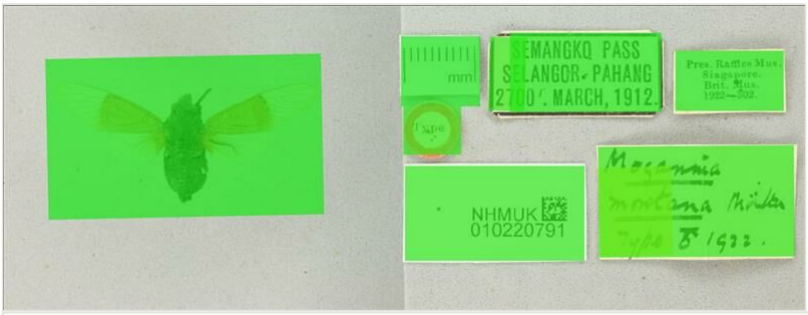

Figure 2.

Visualisation of an openDS FDO within Galaxy, with detected regions of interest (specimen, labels and scale bar) for a pinned insect.